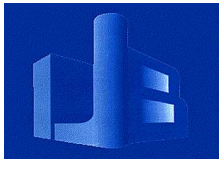
ULB MLG

Use of Machine Learning in Bioinformatics to Identify Prognostic and Predictive Molecular Signatures in Human Breast Cancer

Benjamin Haibe-Kains

bhaibeka@ulb.ac.be

Université Libre de Bruxelles



Institut Jules Bordet

Table of Contents

Thesis and Contributions

Thesis and Contributions

- Thesis concerns cancer classification using machine learning methods and survival analysis based on gene expression data.
- My contributions :
  - **Methodology** (Chapter 4)
  - Preprocessing methods (Sections 4.2.2 and 4.2.3)
  - **Feature selection** (Section 4.3)
  - Classifier validation on different microarray platforms (Section 4.3.2.1)
  - Time-dependent ROC curve (Section 4.5.4)
  - **Cutoff selection** (Section 4.4.2)

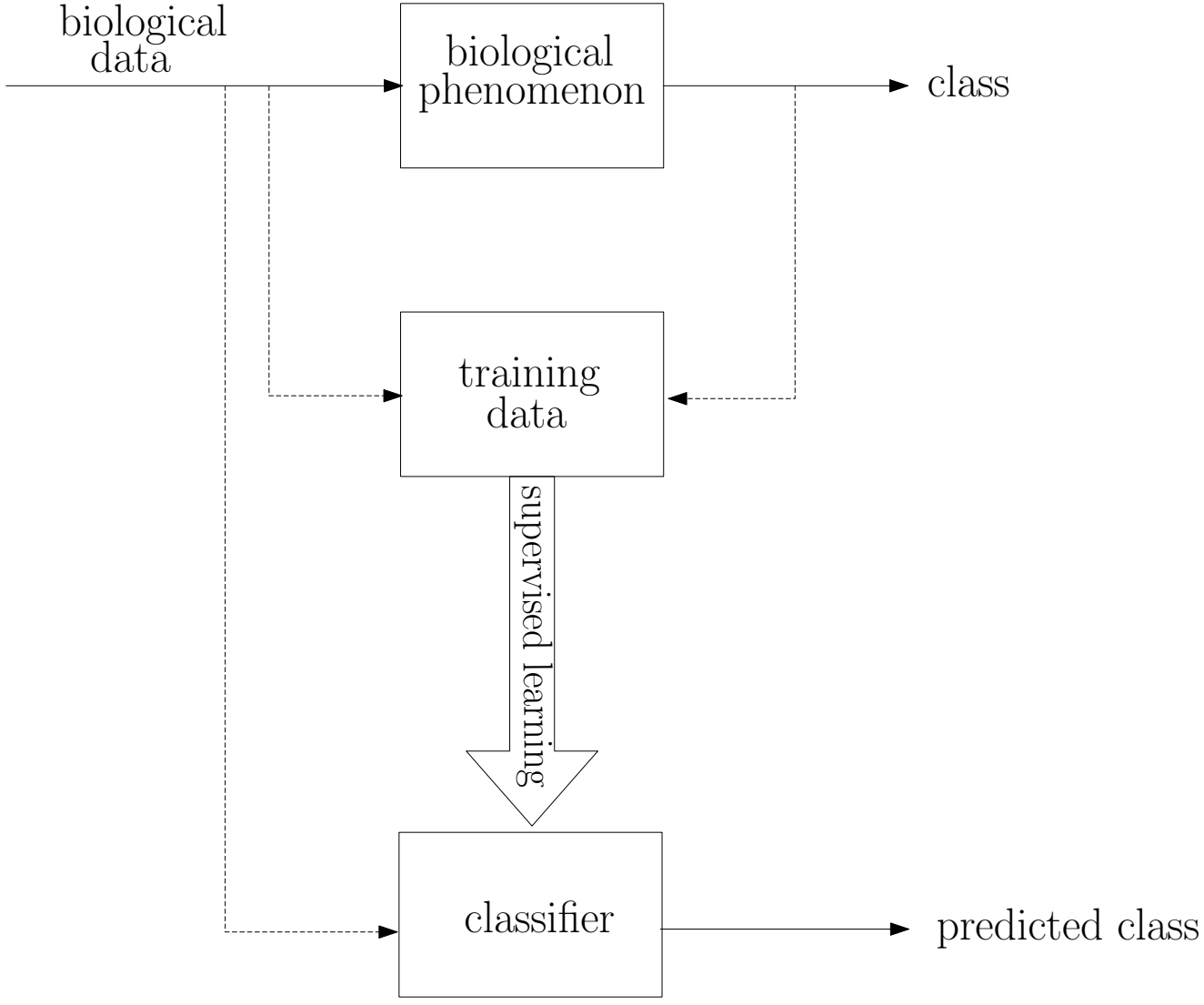
Machine Learning in Bioinformatics

Machine Learning in Bioinformatics

- ***Machine Learning*** *is a field of artificial intelligence related to data mining and statistics, involving learning from data.*
- ***Bioinformatics*** *is the use of techniques from applied mathematics, informatics, statistics, and computer science to solve biological problems.*

- Increasing use of Machine Learning (ML) methods in Bioinformatics over time. This includes cancer classification, gene regulation networks, protein structure, etc.

Example of ML in Bioinformatics



Breast Cancer

Breast Cancer

- The cancer is a **genetic disease** at the level of the cell and affect different types of organs as **breast**.
- Current classifications have serious limitations. Tumors with similar histological criteria
  - can follow significantly different clinical courses (prognosis)
  - can show different responses to therapy (prediction).
- Importance of studying **multiple genetic alterations** in cancer ➡ use of **microarray technology**.

Breast Cancer and Microarray

- The microarray technology provides the opportunity of correlating **genome-wide** expressions with the cancer evolution with/without therapies.

- It is expected that variations in gene expression patterns in different tumors could provide a **molecular signature** of each tumor, and that the tumors could be classified into subtypes based solely on the difference of expression patterns.

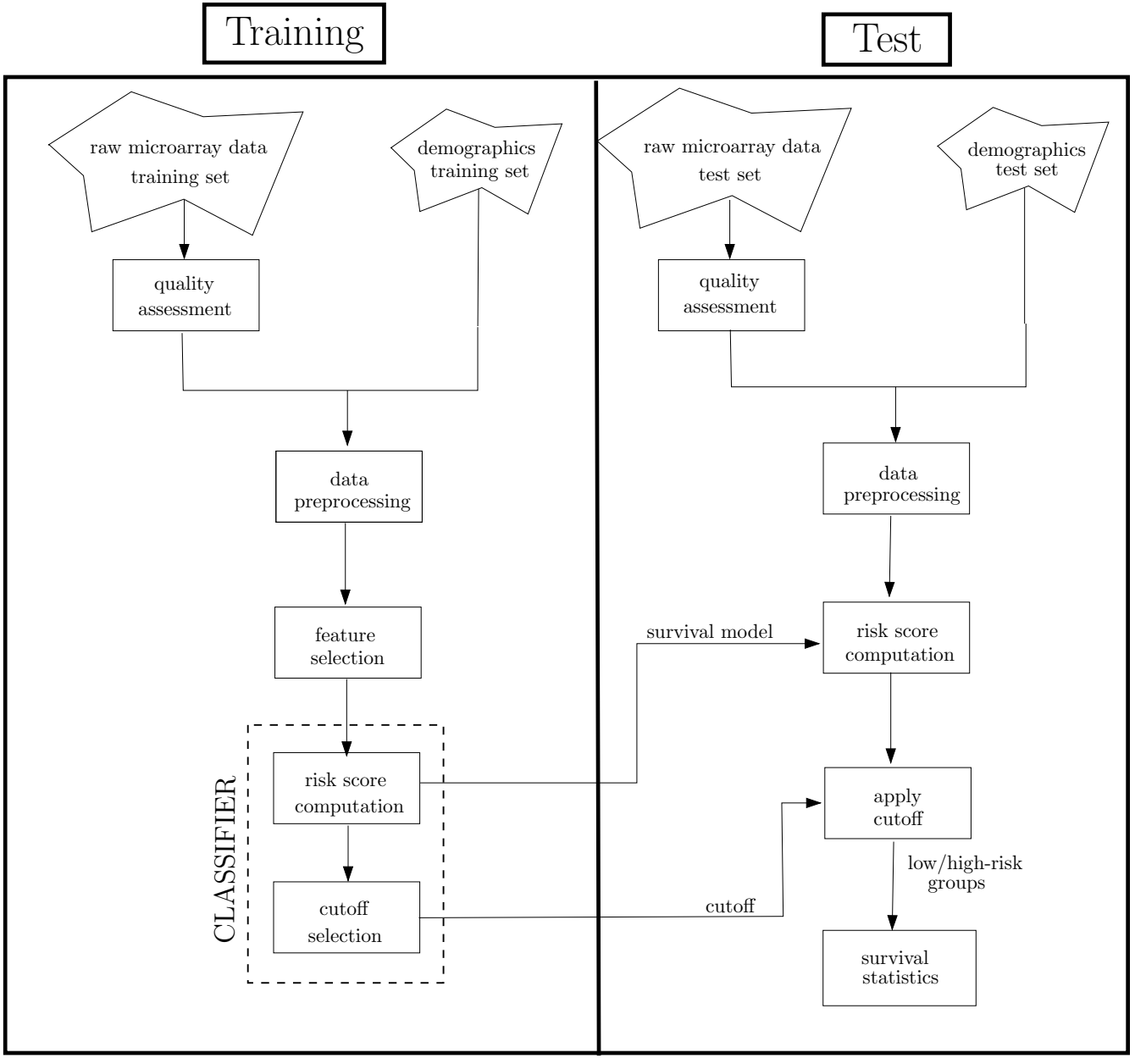
TAMOXIFEN© Resistance

- Joint project with Microarray Unit headed by Dr. Sotiriou.
- Motivations :
  - 40% of patients receiving TAMOXIFEN© will relapse and develop incurable metastatic disease
  - the goal of the data mining analysis is to identify those patients at higher risk of TAMOXIFEN© resistance on the basis of their genetic profile.

Loi, S., Piccart, M., Haibe-Kains, B., Desmedt, C., A.Harris, Bergh, J., Ellis, P., Miller, L., Liu, E., and Sotiriou, C. (2005). *Prediction of early distant relapses on tamoxifen in early-stage breast cancer : a potential tool for adjuvant aromatase inhibitor tailoring*. In Proceedings of ASCO Meeting, abstract 509.

Methods and Results

Methodology



Modeling Survival Data

- We use the semiparametric regression model proposed in [Cox, 1972] to estimate hazard functions given a dataset.
  - no assumption about the probability distribution of survival times (**proportional hazards model**)
  - efficient estimation method (**maximum partial likelihood**).
- Hazard function :

$$h_i(t) = \underbrace{\lambda_0(t)}_{\text{baseline hazard function}} \exp \underbrace{(\beta_1 x_{1i} + \cdots + \beta_n x_{ni})}_{\text{risk score}}$$

where β_j are the coefficients to estimate and x_{ji} are the gene expressions

Maximum Partial Likelihood

- Basic model : $h_i(t) = \lambda_0(t) \exp(\beta_1 x_{1i} + \cdots + \beta_n x_{ni})$
- Proportional hazards model :

$$\frac{h_i(t)}{h_j(t)} = \exp\{\beta_1(x_{1i} - x_{1j}) + \cdots + \beta_n(x_{ni} - x_{nj})\}$$

➥ $\lambda_0(t)$ cancels out.

- Partial likelihood function : $PL = \prod_{i=1}^{N} L_i$

where $L_i = \dfrac{h_i(t_j)}{\sum_{k \in at\ risk} h_k(t_j)}$.

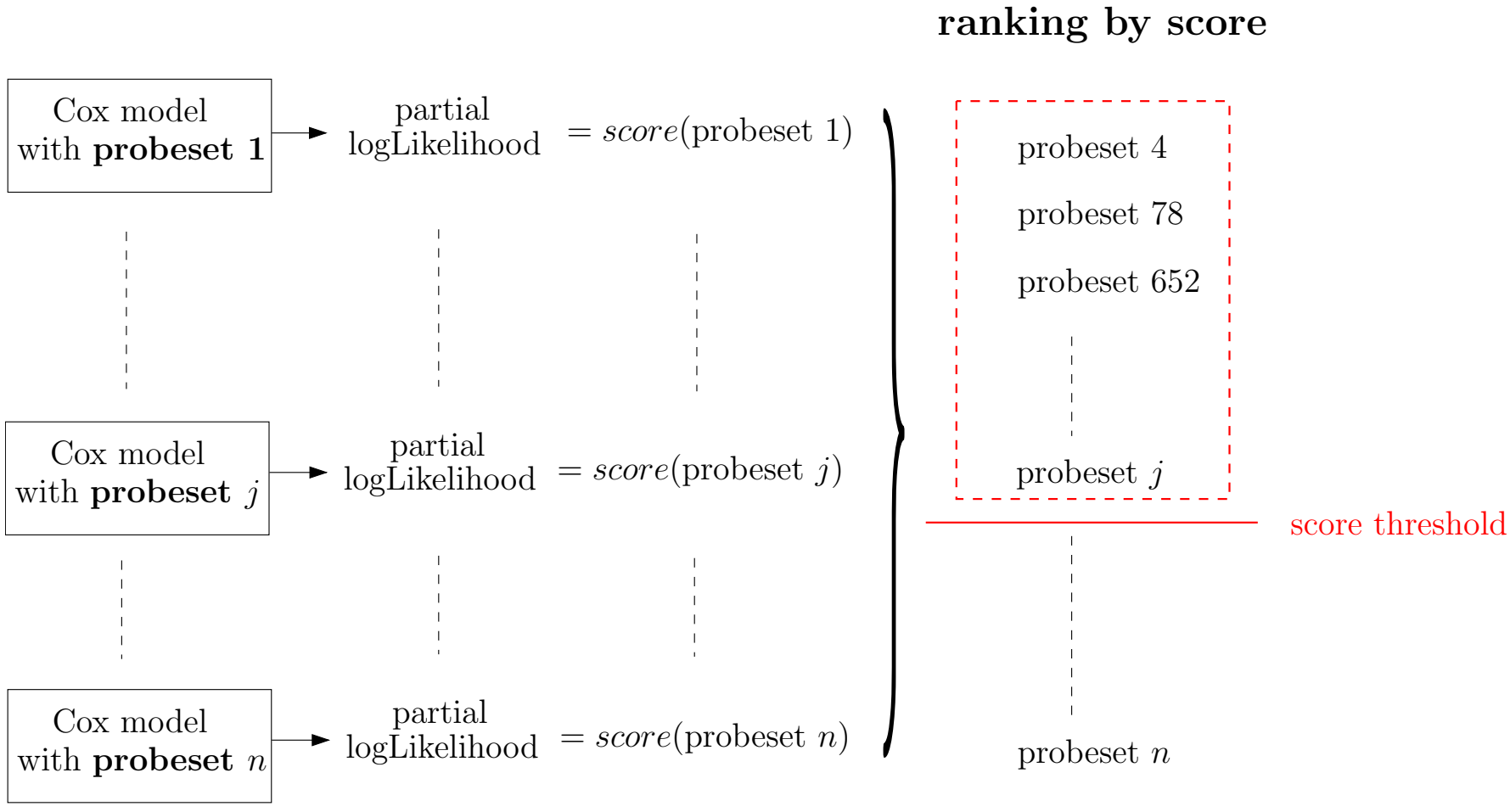
Feature Selection

- Microarray data characteristics
  - **high feature/sample ratio** : thousands of probesets (up to 50,000) and hundreds of patients
  - **highly correlated features** : co-regulation of many genes.
- Potential benefits of feature selection

  [Guyon and Elisseeff, 2003]
  - facilitating data visualization and understanding
  - reducing measurement and storage requirements
  - reducing training and computation time
  - defying the curse of dimensionality to improve prediction performance.

Variable Ranking

- Variable ranking based on univariate Cox model.



Feature Construction

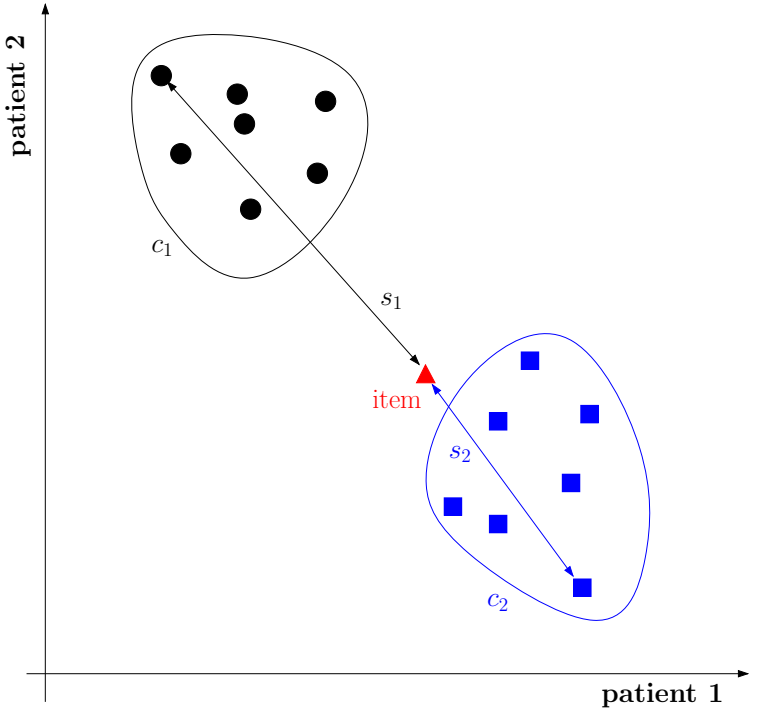
- Hierarchical clustering to compute cluster centroids of **highly correlated probesets** previously selected by variable ranking.
- Test different number of clusters.
- For each number of clusters
  - compute cluster centroids
  - estimation of performance by 10-fold cross-validation using multivariate Cox model.

➥ best number of cluster is 2.

Hierarchical Clustering

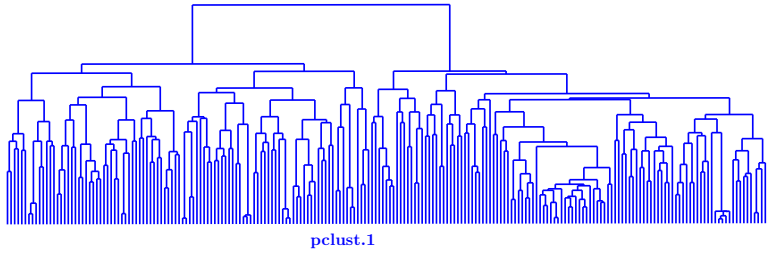
- Common clustering method [Hartigan, 1975; Eisen et al., 1998].
- Organizing probesets in a hierarchical binary tree (**dendrogram**) based on their **degree of similarity**.

- Metric of similarity : *uncentered Pearson correlation*.
- Linkage : *complete linkage*.

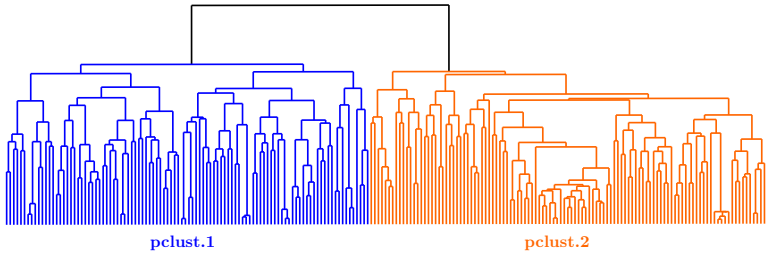


Number of Clusters

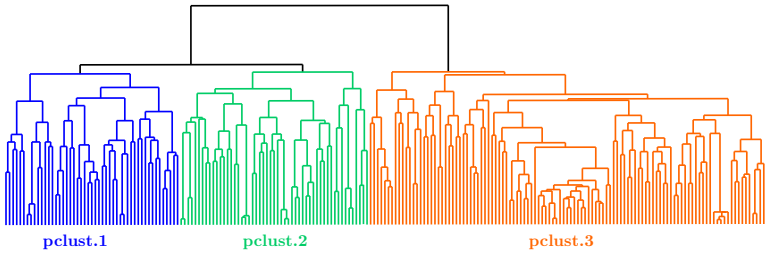
number of cluster $= 1$



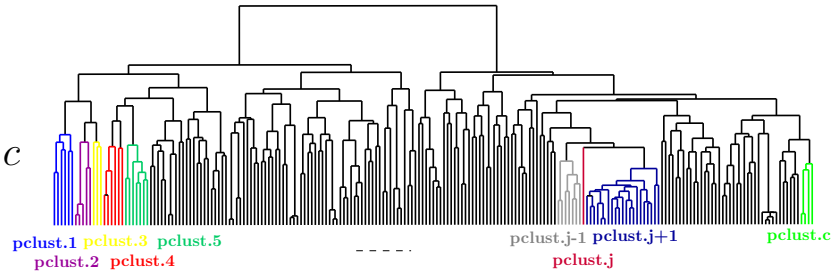
number of cluster $= 2$



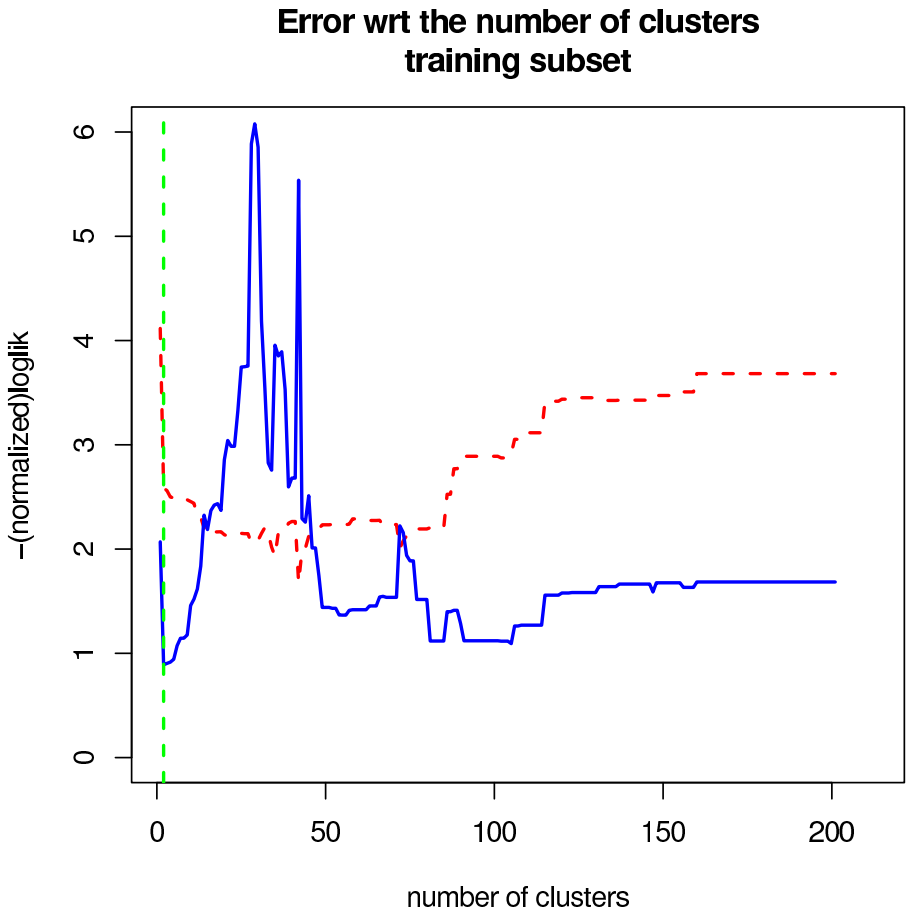
number of cluster $= 3$



number of cluster $= c$



Performance Estimate



- -(normalized) logLikelihood is the error measure.
- dashed line is the training error.
- solid line is the test error.
- vertical dashed line is the best number of clusters (2).

Final Model

- Fitting a multivariate Cox model using all the training set (OXFT).
- Risk score computation :

$$rs_i = \sum_{j=1}^{C} \hat{\beta}_j c_{ji}$$

where rs_i is the risk score of patient i, $\hat{\beta}_j$ are the estimated coefficients, C is the number of clusters and c_{ji} are the cluster centroids.

Survival Statistics for 2 groups

There exist several ways to assess difference in survival between 2 groups

- **Kaplan-Meier estimator** and **Logrank test** :
  - KM method estimates survivor function such that

$$\widehat{S}(t) = \prod_{j:t_j \leq t} [1 - \frac{\overbrace{d_j}^{\#\text{ death at time } t_j}}{\underbrace{n_j}_{\#\text{ at risk at time } t_j}}].$$

  - logrank method tests $H_0 : S_1(t) = S_2(t)\ \forall t \geq 0$.
- **Hazard ratio** (HR) : relative hazard[a] between 2 groups using Cox model with one dummy variable ($G = 0/1$ for low and high-risk groups).

[a] HR = 1 means no difference in survival between 2 groups.

Cutoff Selection

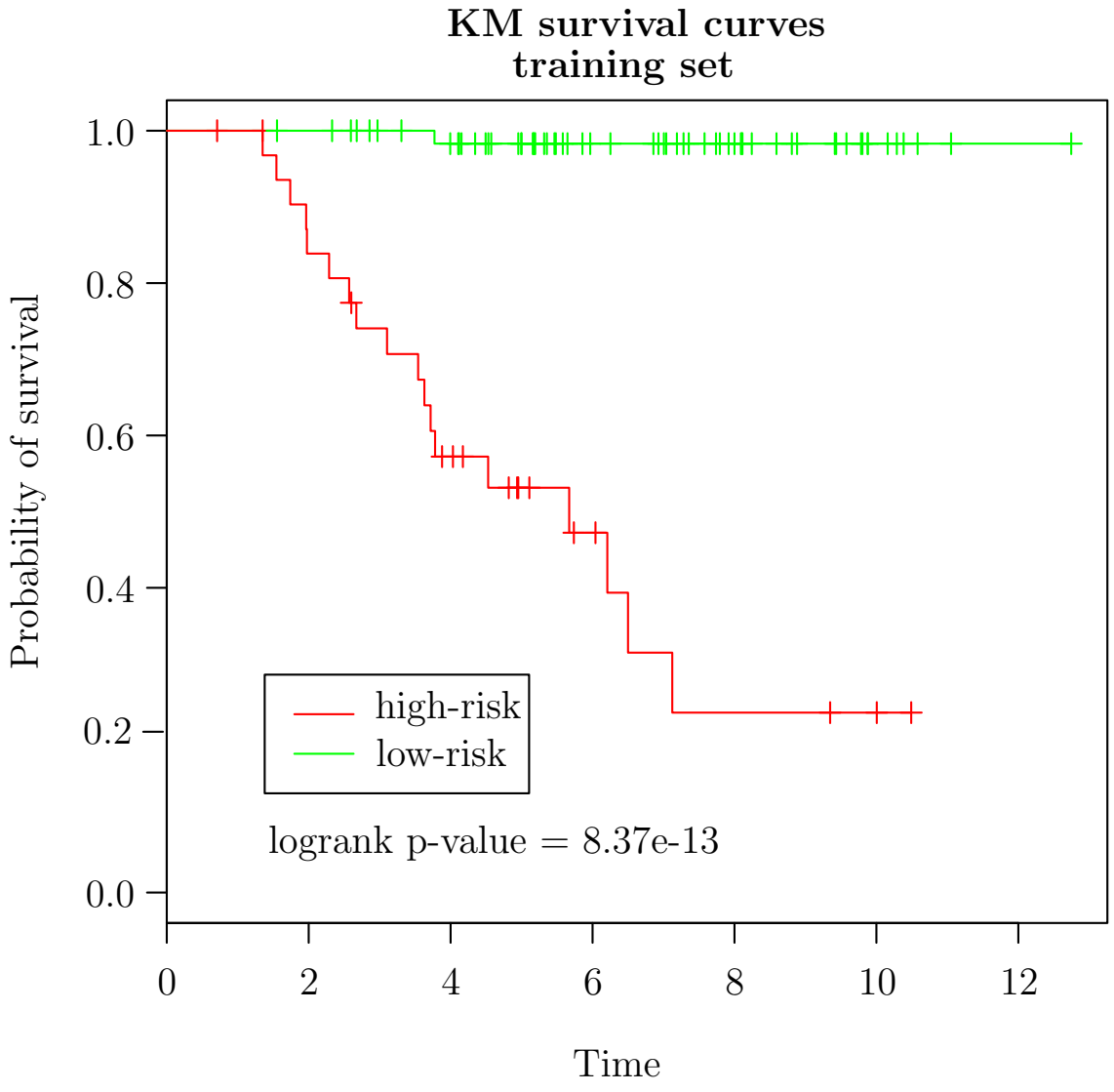
Algorithm :

1. Consider only the training set.
2. Keep only cutoffs which leave at least 25% of patients in one group.
3. Keep only cutoffs which have not 1 in 95% CI for HR.
4. Select the cutoff with the lowest proportion of DMFS[a] at 3 years in the low-risk group and the highest HR.

Results : cutoff = 0.93 with hazard ratio = 60.42 (95% CI [7.99, 456,59]), proportion of DMFS = 0% and 67.6% (67/99) of patients in low-risk group.

[a] Distant Metastasis Free Survival.

KM Survival Curves

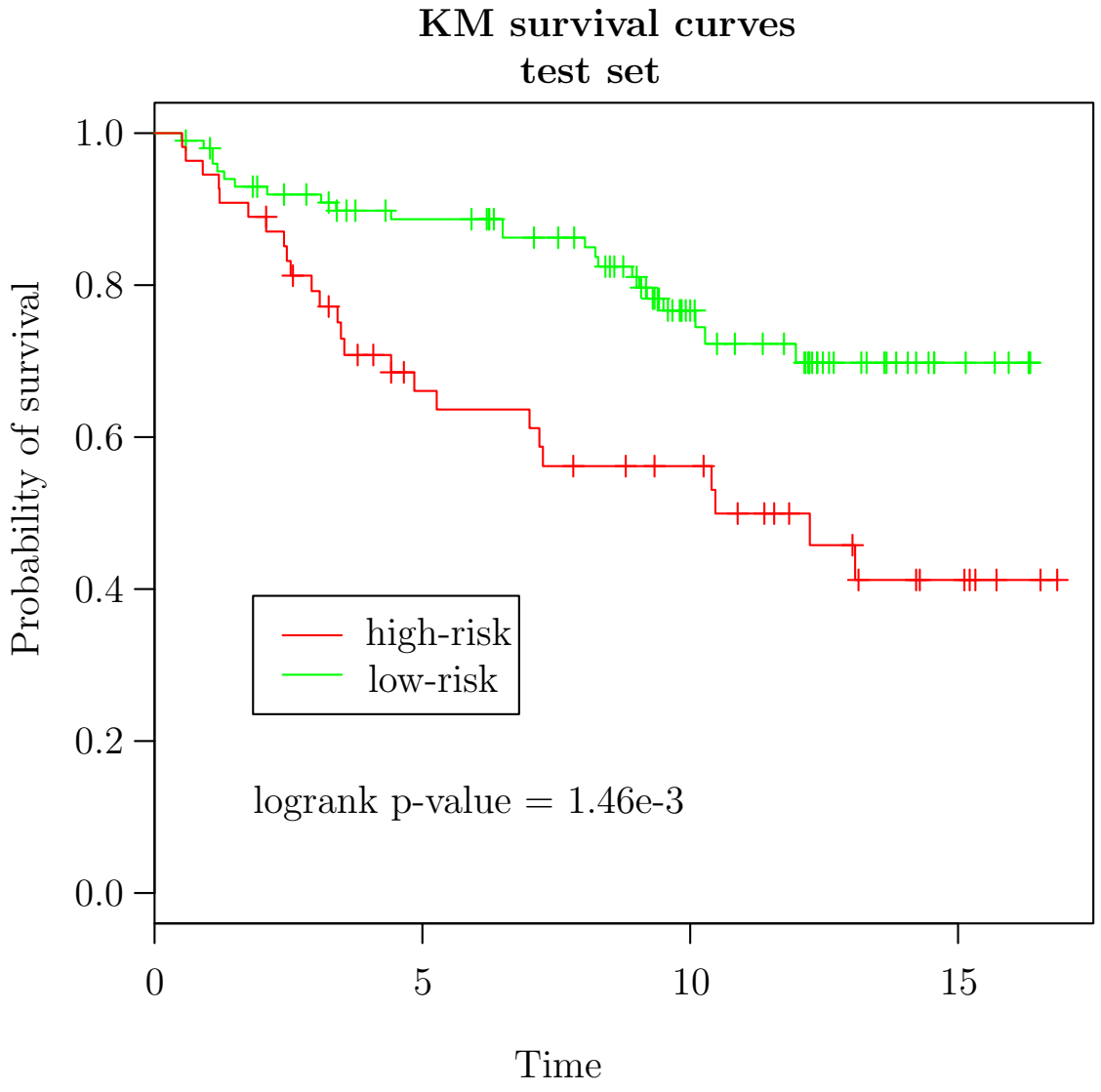


Validation on Test Set

- Using the same formula to compute the risk score and the same cutoff in the **independent test set** (KIT and GUYT), we have
  - hazard ratio = 2.44 (95% CI [1.38, 4.31])
  - proportion of DMFS = 8% in the low-risk group
  - 64.7% (101/156) of patients in low-risk group.

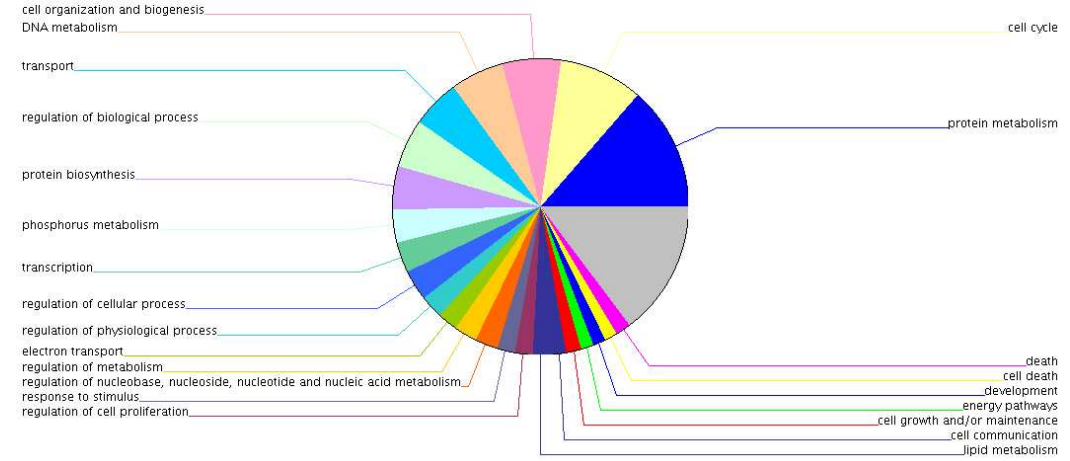
➥ **significant difference in survival** between low and high-risk groups in the **test set**.

KM Survival Curves

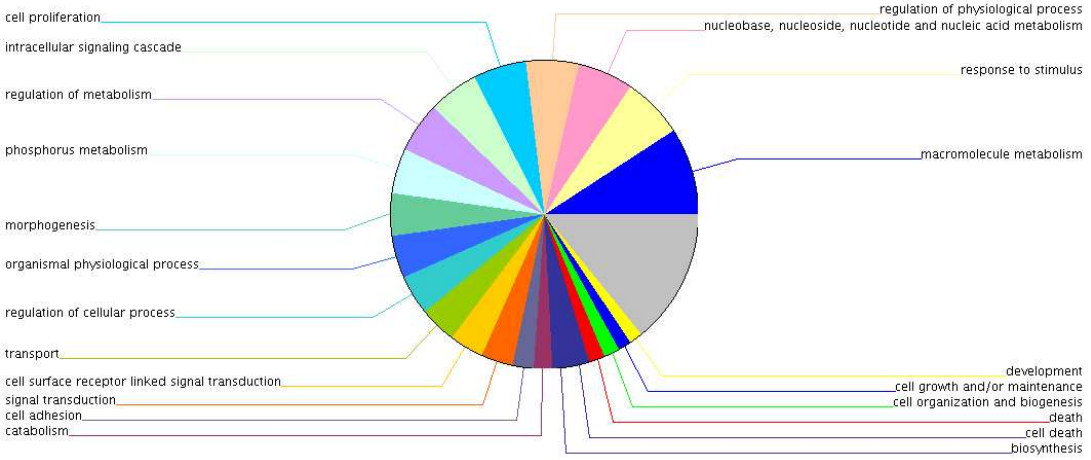


Gene Ontology

cluster 1 :



cluster 2 :



Lacroix, M., Haibe-Kains, B., Laes, J. F., Hennuy, B., Lallemand, F., Gonze, I., Cardoso, F., Piccart, M., Leclercq, G., and Sotiriou, C. (2004). *Gene regulation by phorbol 12-myristate 13-acetate (PMA) in two highly different breast cancer cell lines*. Oncology Report, 12(4):701-707.

Conclusion

Conclusion

- New methodology covering the whole range of microarray analysis
  - machine learning methods (e.g. feature selection, cross-validation)
  - well-established survival methods (e.g. Cox model, survival statistics).
- Final classifier remaining biologically interpretable.
- Potential robustness for validation on different microarray platforms.
- Successful test on real data dealing with the TAMOXIFEN© resistance of breast cancer patients.

Future Works

- Impact and implementation of preprocessing methods on specific computer architectures (e.g. computers cluster).
- Study of variance of variable ranking.
- Study of penalized Cox model [Tibshirani, 1997; Gui and Li, 2004]
- Alternative methods for feature construction.
- Study of the classifier robustness with the loss of one or more probesets (validation on different microarray platforms).

Future Works(2)

- Multiple random validation strategy [Michiels et al., 2005].
- Comparison with binary classification techniques [Dudoit et al., 2002; Haibe-Kains, 2004].
- Use of Gene Ontology to infer biological knowledge.
- Comparison with traditional histological criteria.
- Comparison with other molecular signatures [Paik et al., 2004; Ma et al., 2004].

Links

- Personal homepage :
  `http://www.ulb.ac.be/di/map/bhaibeka/.`
- Microarray Unit :
  `http://www.bordet.be/servmed/array/.`
- Machine Learning Group :
  `http://www.ulb.ac.be/di/mlg.`
- DEA/DES in Bioinformatics :
  `http://www.bioinfomaster.ulb.ac.be/.`

Thanks for your attention

Benjamin Haibe-Kains

References

Bolstad, B. M., Irizarry, R. A., Astrand, M., and TP, T. S. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Chapman and Hall, second edition edition.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society Series B*, 34:187–220.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863–14868.

Gui, J. and Li, H. (2004). Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Center for Bioinformatics and Molecular Biostatistic, paper L1Cox*. http://repositories.cdlib.org/cbmb/L1Cox.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Haibe-Kains, B. (2004). Breast cancer diagnosis using microarray. Master's thesis, ULB.

Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley.

Ma, X. J., Wang, Z., Ryan, P. D., Isakoff, S. J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., Tuggle, J. T., Tran, Y., Tran, D., Tassin, A., Amon, P., Wang, W., Wang, W., Enright, E., Stecker, K., Estepa-Sabal, E., Smith, B., Younger, J., Balis, U., Michaelson, J., Bhan, A., Habion, K., Baer, T. M., Brugge, J., Haber, D. A., Erlander, M. G., and Sgroi, D. S. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell*, 5:607–616.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple radom validation strategy. *Lancet*, 365:488–492.

Paik, S., Shak, S., Tang, G., Kim, C., Bakker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J., and Wolmark, N. (2004). A multigene assay to predict recurrence

of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medecine*, (351):2817–2826.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medecine*, 16:385–395.

Appendix

Microarray Unit

- Laboratory of the Institut Jules Bordet (IJB).
- 6 researchers (funds coming from IJB, Télévie, FNRS, etc.).
- Numerous running projects concerning the breast cancer :
  - Appearance of distant metastases
  - Response to therapies
  - Refinement of histological criteria by microarray, etc.
- Scientific collaborations :
  - Singapore Lab (Lence Miller and Edison)
  - Swiss Institute of Bioinformatics, etc.

Microarray Unit(2)

- Biological and medical facilities :
  - AFFYMETRIX© and cDNA microarray platforms
  - all the kits necessary to check the quality of the biological samples (AGILENT©) and to perform microarray experiments
  - access to the tumor bank of IJB.
- Computing facilities :
  - 2 workstations (P4 3.6 GHz, 4 Go RAM)
  - access to LIT5 cluster.
- Website : `http://www.bordet.be/servmed/array/`.

Members of Microarray Unit

Machine Learning Group

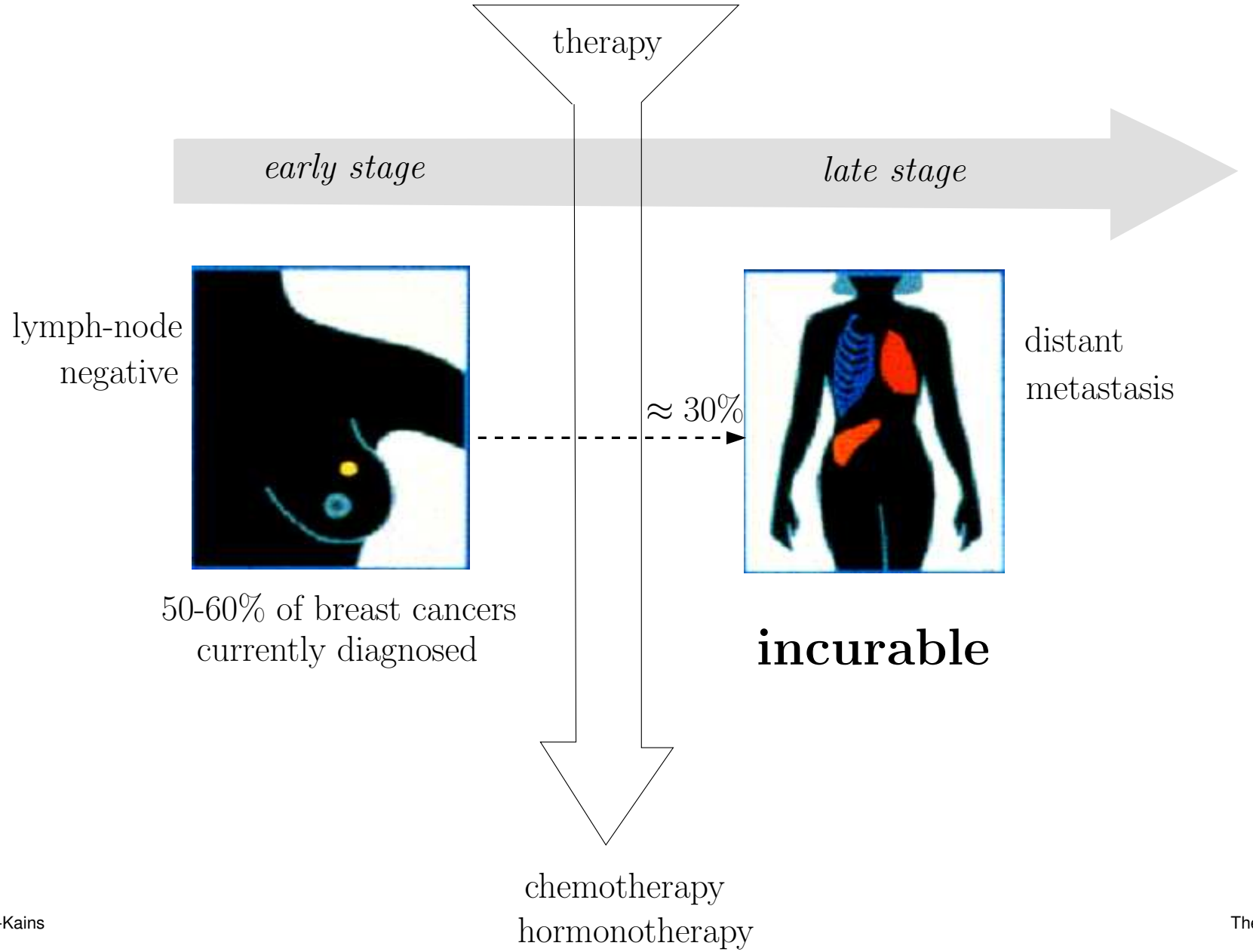
- Research group of the Université Libre de Bruxelles (ULB).
- 7 researchers (funds coming from ULB, ARC, European Community, etc.).
- Research topics :
  - Local learning, Classification, Computational statistics, Data mining, Regression, Time series prediction, Sensor networks, Bioinformatics.
- Scientific collaborations in ULB :
  - IRIDIA (Sciences Appliquées), Physiologie Moléculaire de la Cellule (IBMM), Microarray Unit (IJB), Service d'Anesthésie (ERASME), etc.

Machine Learning Group(2)

- Scientific collaborations outside ULB :
  - UCL Machine Learning Group (B), Politecnico di Milano (I), Università del Sannio (I), George Mason University (US), etc.
- Computing facilities :
  - LIT5 cluster (16 x P4 3.4 GHz, 16 x 2 Go RAM)
  - LEGO Robotics Lab.
- Website : `http://www.ulb.ac.be/di/mlg`.

Members of MLG

Breast Cancer Practice



Breast Cancer Therapy

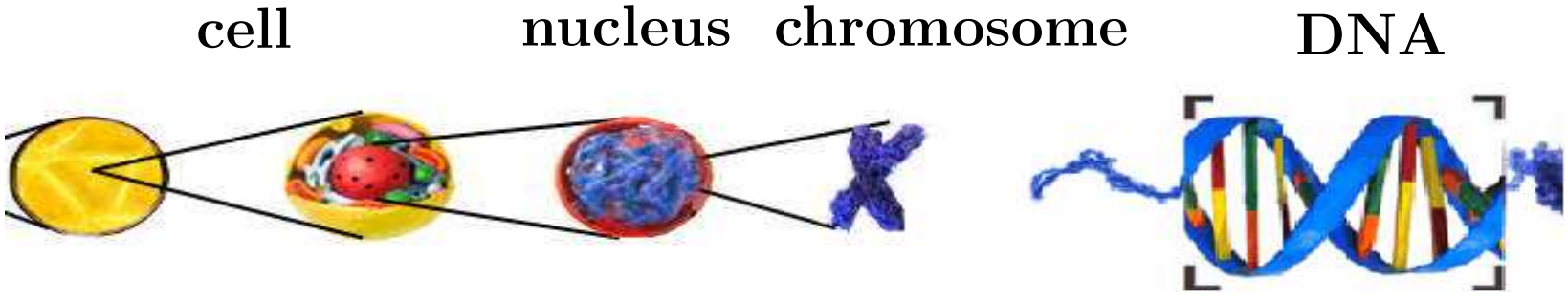
- Chemotherapy or hormonotherapy reduces the risk of distant metastasis by $2-12\%$
- However
  - $\approx 70\%$ of patients receiving this treatment would have survived without it
  - these therapies frequently have **toxic side effects**.
- Classification of cancers must be accurate in order to give the correct treatment and so increase the chance of survival for the patient.

TransBIG Consortium

- Joint project with Microarray Unit headed by Dr. Sotiriou.
- Motivations :
  - current risk evaluation of early breast tumors (see St Galln, NIH and NPI) fails to classify correctly the tumors. It results :
    - unnecessary therapies
    - toxic side effects
    - waste of money
  - the goal of the data mining analysis is to identify those patients at higher risk of distant metastases appearance on the basis of their genetic profile.

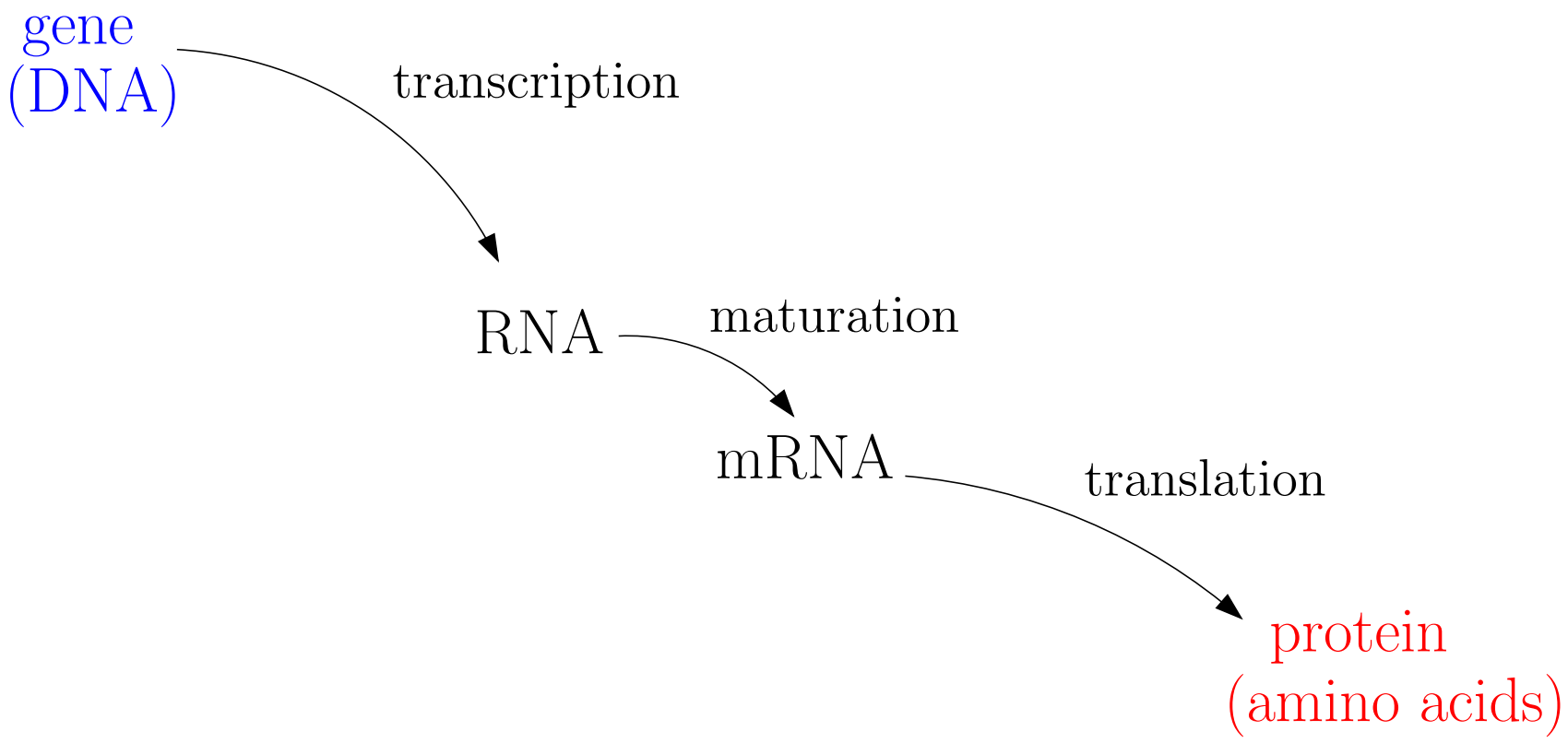
Genetics : basics

Cell to DNA :



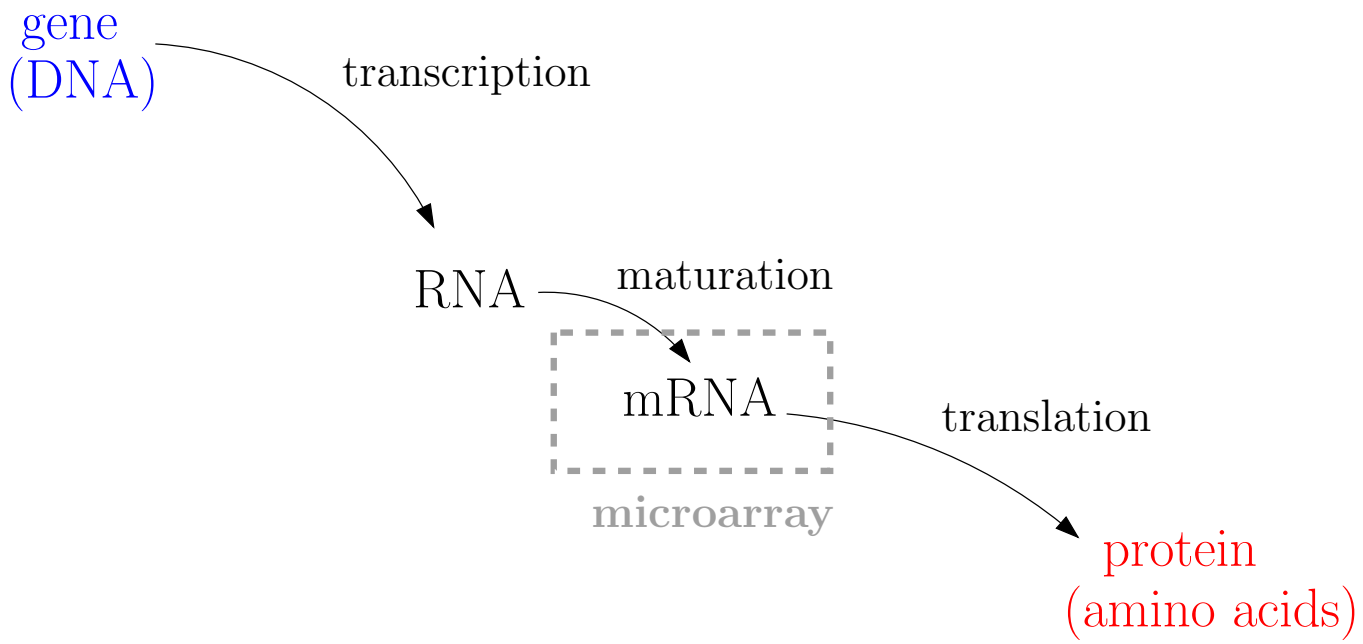
Genetics : basics(2)

DNA to protein :



Microarray Technology

- As we will see, the microarray technology allows us to study the **genetic profile** of breast tumors.
- Microarray works at the mRNA level :



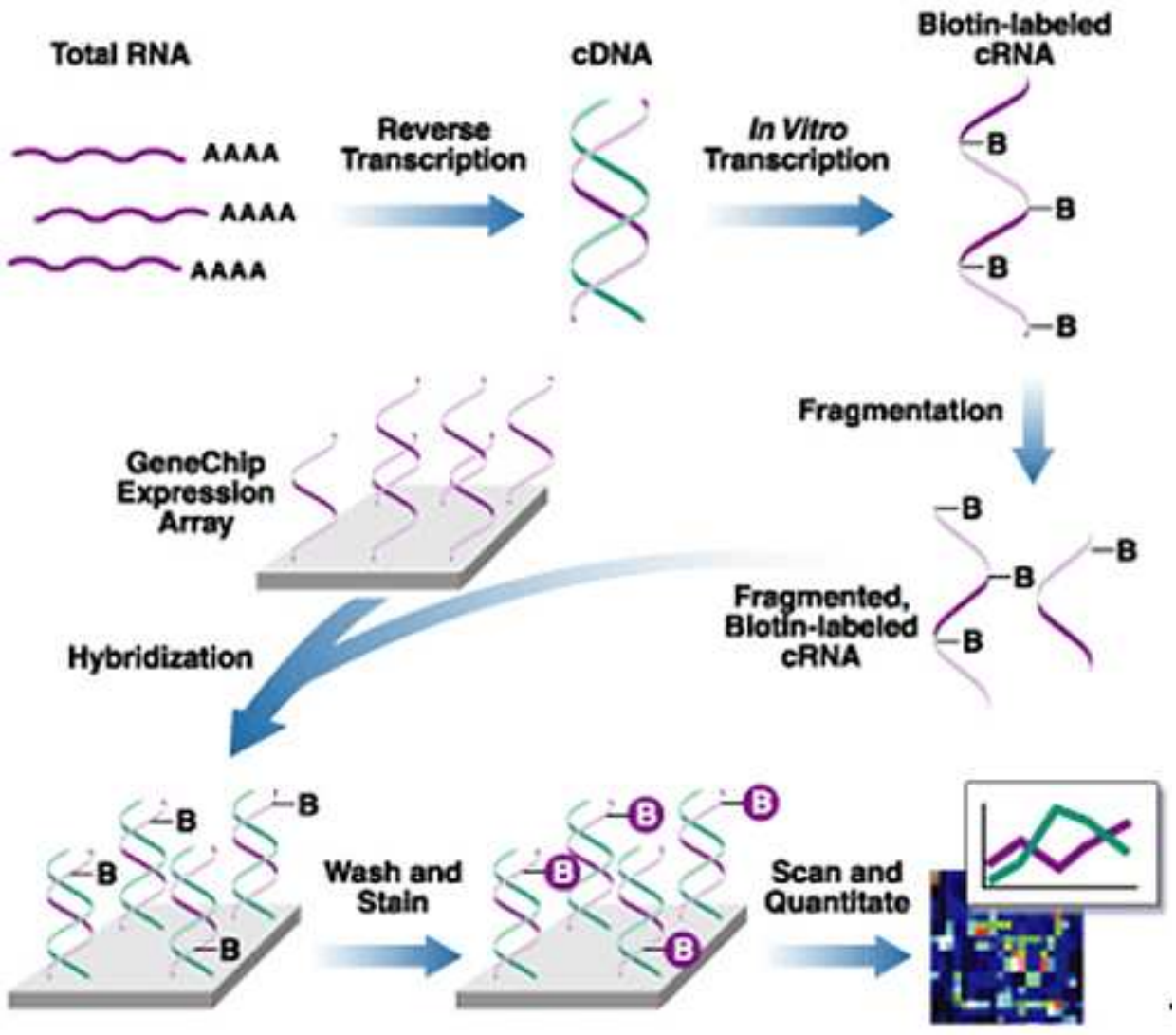
Microarray Technology Description

A *microarray* is composed of

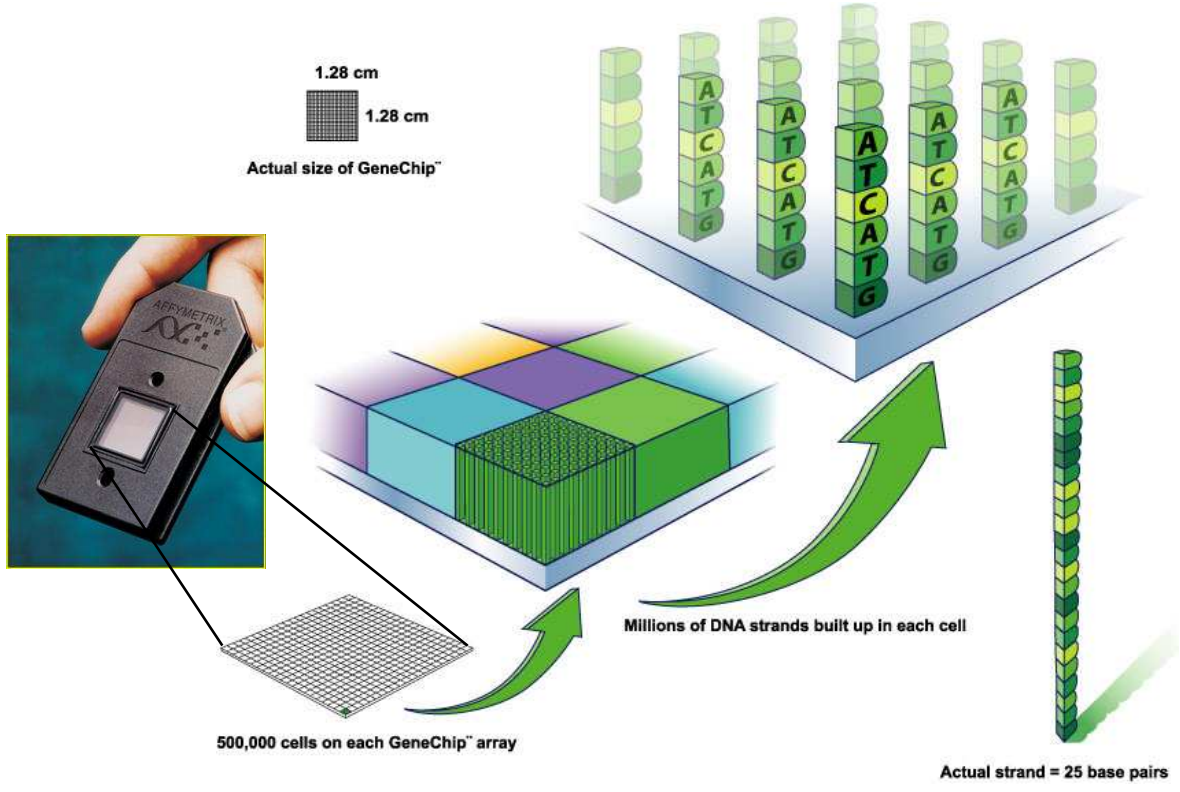
- DNA fragments fixed on a solid support
- ordered position of probes
- principle of hybridization to a specific probe of complementary sequence
- radioactive labeling

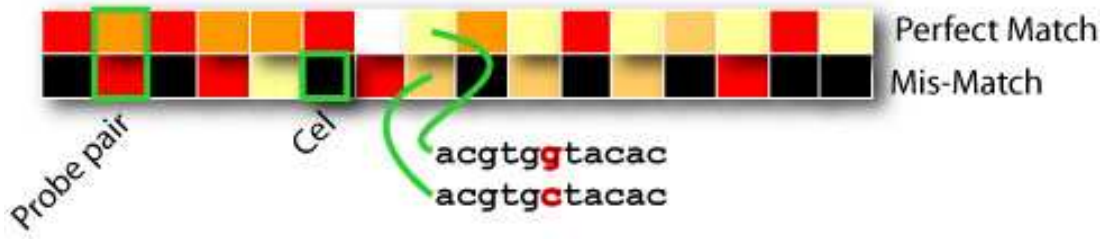
➥ simultaneous detection of thousands of sequences in parallel

AFFYMETRIX© Design

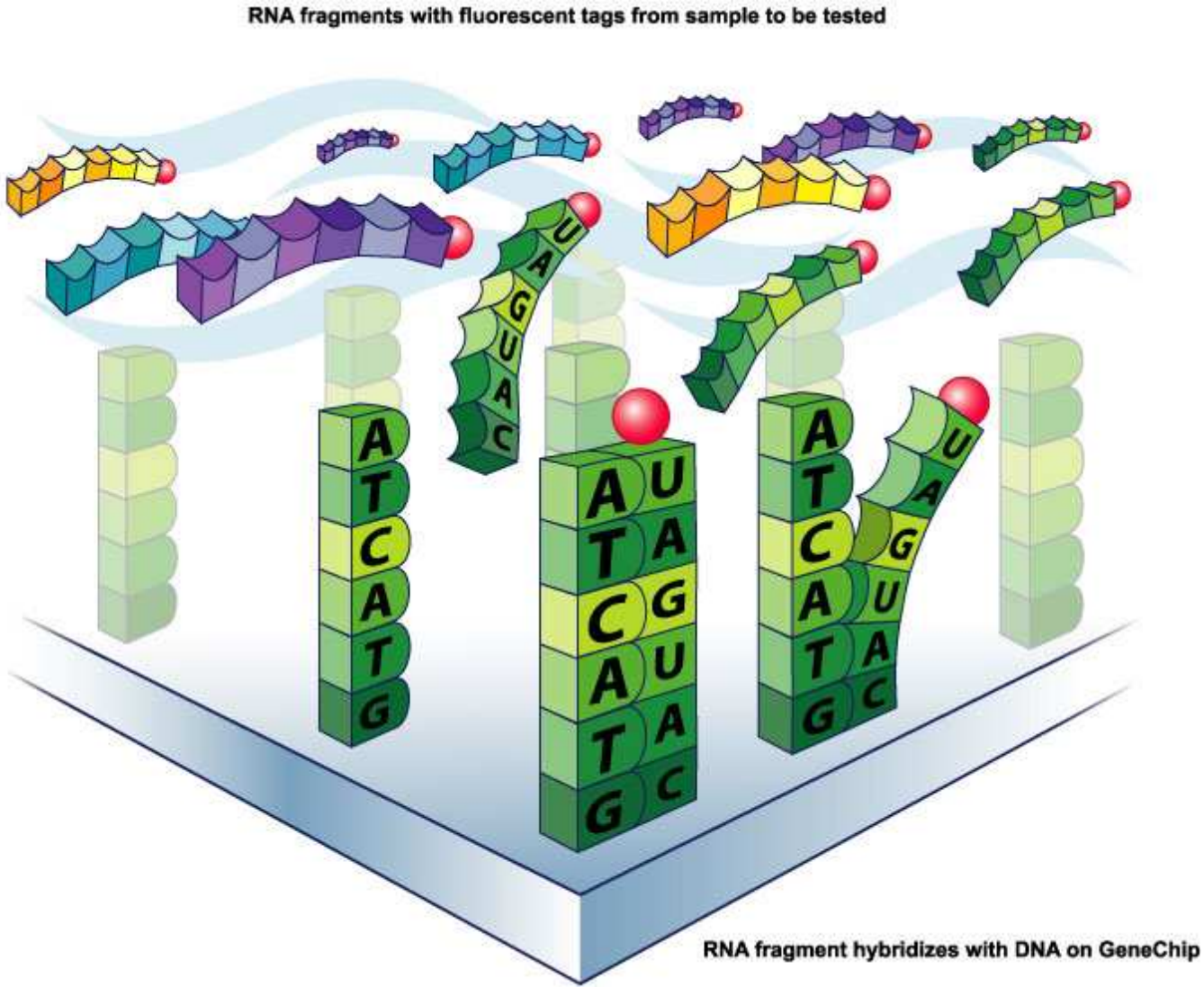


AFFYMETRIX© GeneChip Structure



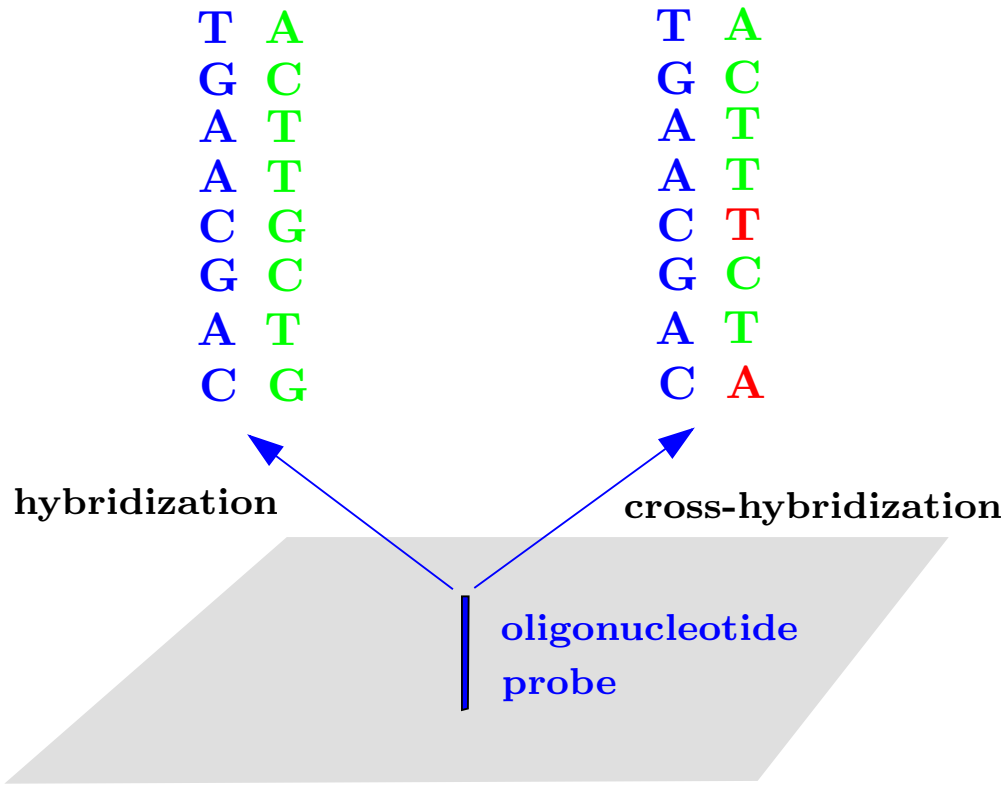


AFFYMETRIX© Hybridization

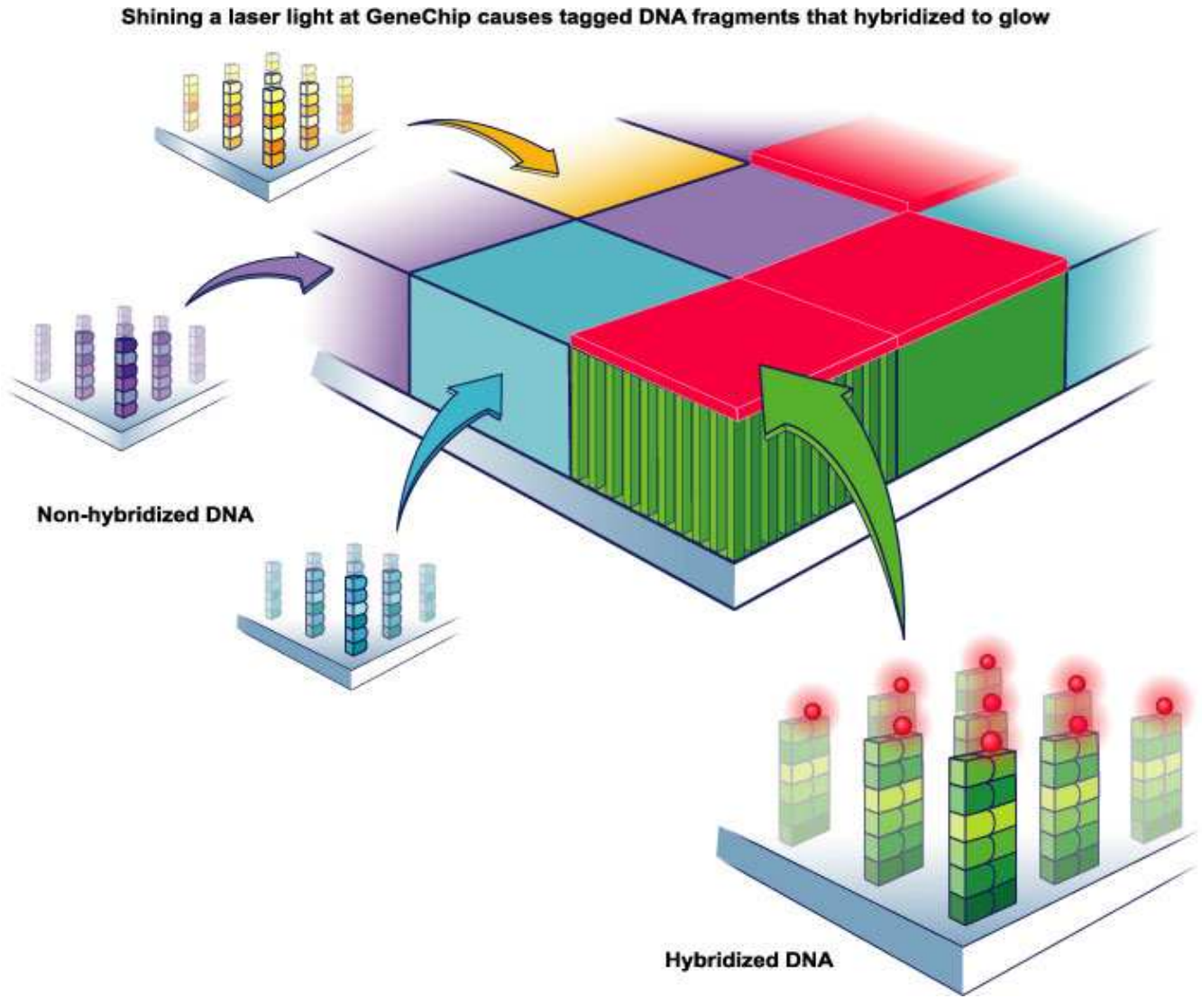


AFFYMETRIX© Cross-Hybridization

- The process of 2 complementary DNA strands binding is called *hybridization*.
- Ideally, an oligonucleotide probe will only bind to the DNA sequence for which it was designed and to which it is complementary.
- However, many DNA sequences are similar to one another and can bind to other probes on the array.
- This phenomenon is called *cross-hybridization*.

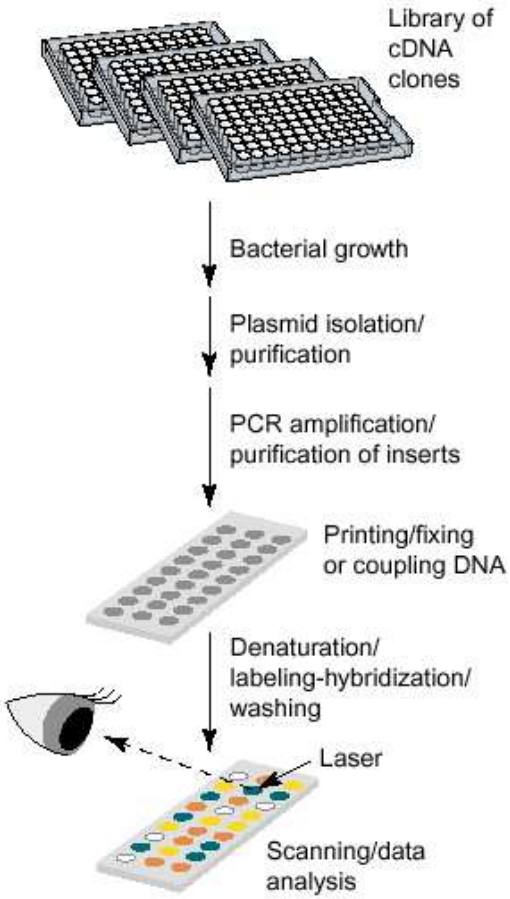


AFFYMETRIX© Detection

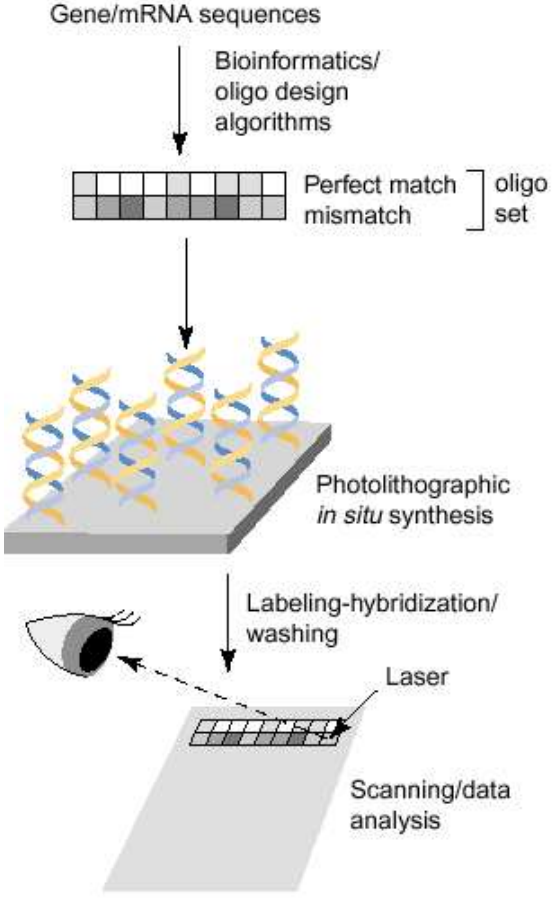


Microarray Comparison

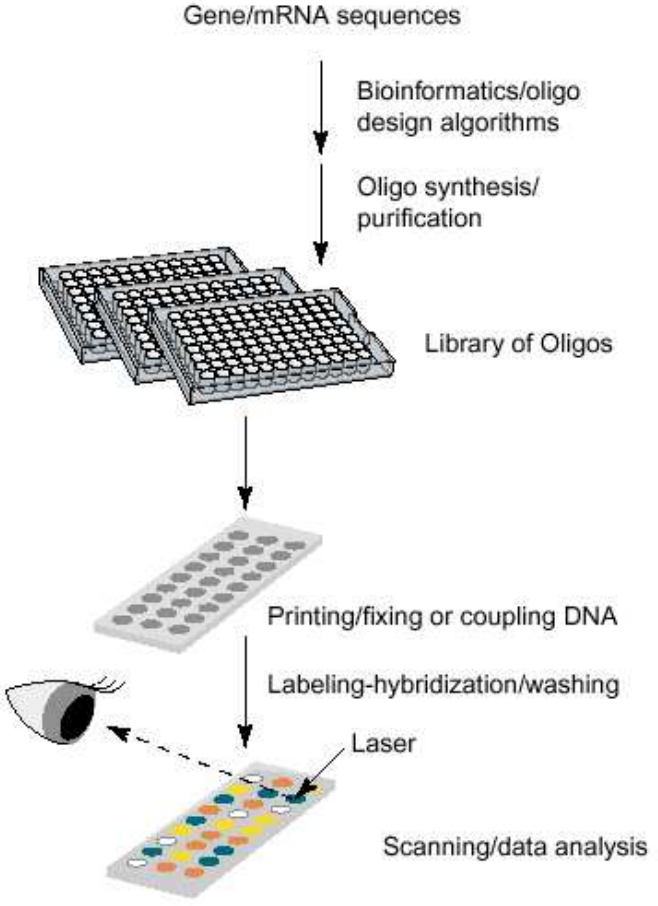
cDNA



short oligonucleotide



long oligonucleotide



AFFYMETRIX© +

AFFYMETRIX© advantages :

- commercially available for several years (strong manufacturing)
- large number of published studies (generally accepted method)
- no reference sample $\rightarrow$ possible comparison between studies

AFFYMETRIX© -

AFFYMETRIX© disadvantages :

- cost of the devices and the chips (but easy use)
- changes in probe design is hard (but new program permits to create his own design)
- short oligos $\rightarrow$ several oligos per gene, specificity/sensitivity trade-off (complex methods to get gene expression)

Data Preprocessing

- Use of RMA separately for each population
  - problem of computer resources
  - data management
- Population correction
  - remove source of variability related to the origin of samples

➥ 44,928 corrected probesets

- Prefiltering based on detection calls (use of MM information)

➥ 32,139 corrected probesets.

Expression Quantification

For each probe set, **summarization** of the probe level data (11-20 PM and MM pairs) into a single expression measure

RMA procedure

- use only PM and ignore MM
- adjust for background on the raw intensity scale
- carry out **quantile** normalization [Bolstad et al., 2003] of $PM - \hat{BG}$ and call the result $n(PM - \hat{BG})$
- take log2 of normalized background adjusted PM
- carry out a **medianpolish** of the quantities $log_2\, n(PM - \hat{BG})$.

Maximum Partial Likelihood

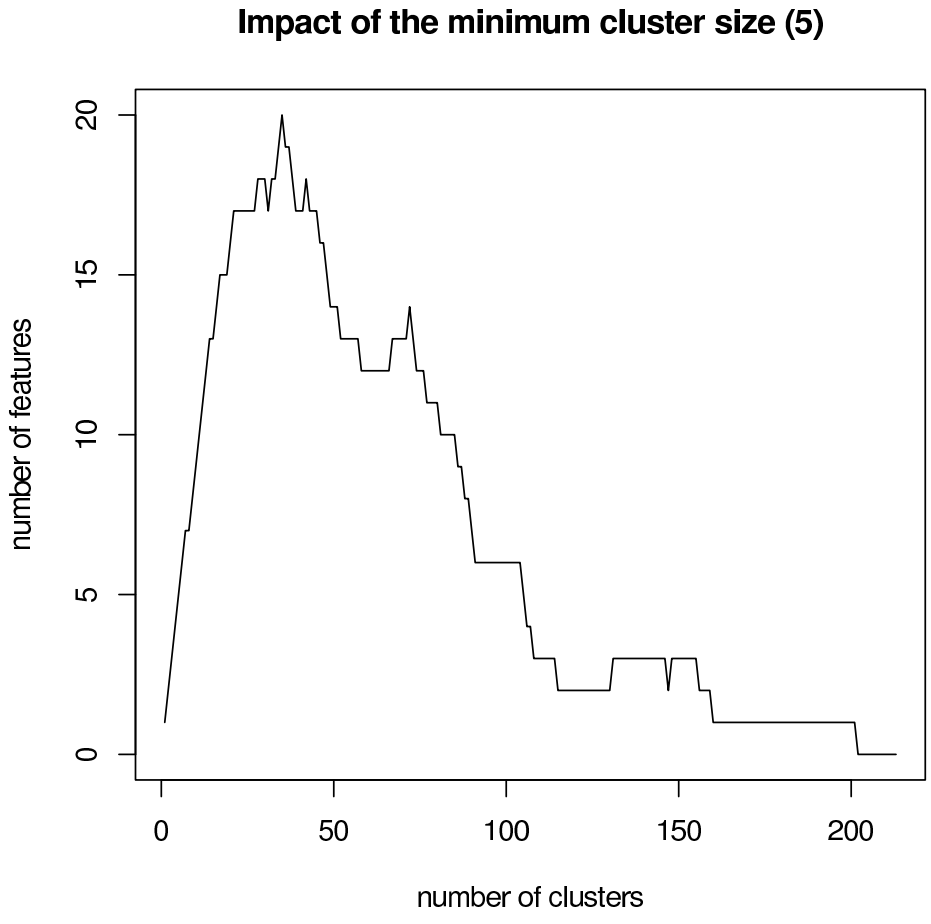
- Logarithm of the partial likelihood :

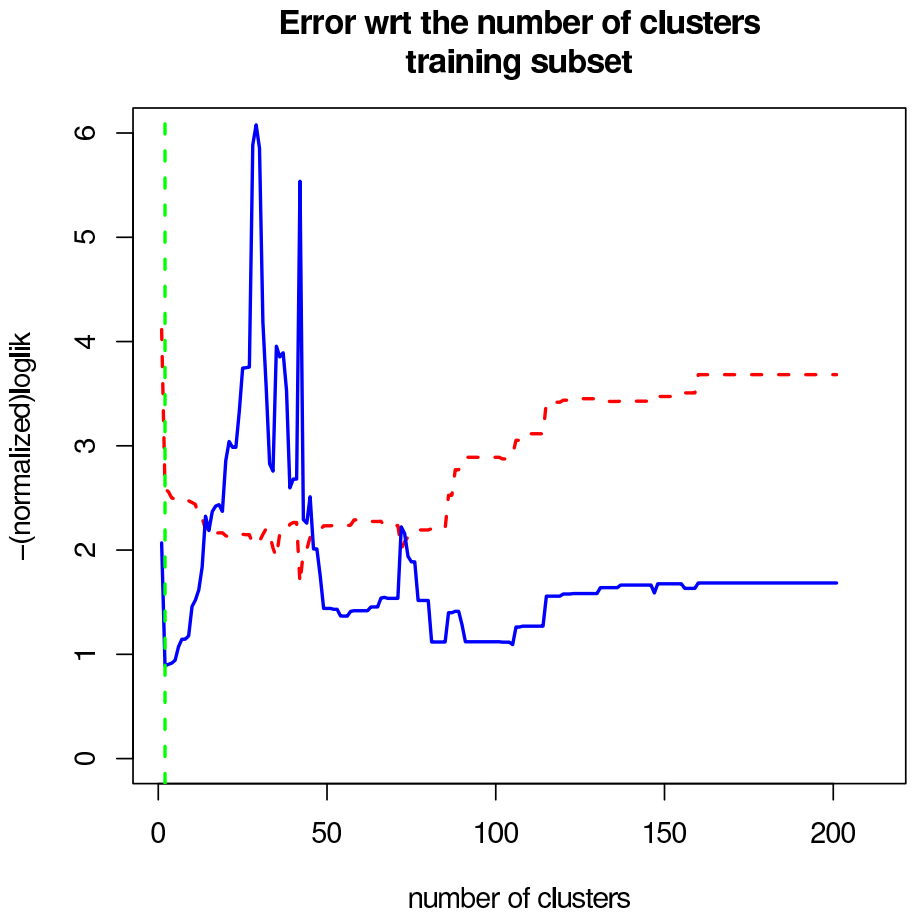
$$\ln PL = \sum_{i=1}^{N} \delta_i \left[\boldsymbol{\beta} \mathbf{x}_i - \ln \left(\sum_{j=1}^{N} y_{ij} e^{\boldsymbol{\beta} \mathbf{x}_j} \right) \right]$$

  where $y_{ij} = 1$ if $t_j \geq t_i$ and $y_{ij} = 0$ if $t_j < t_i$.

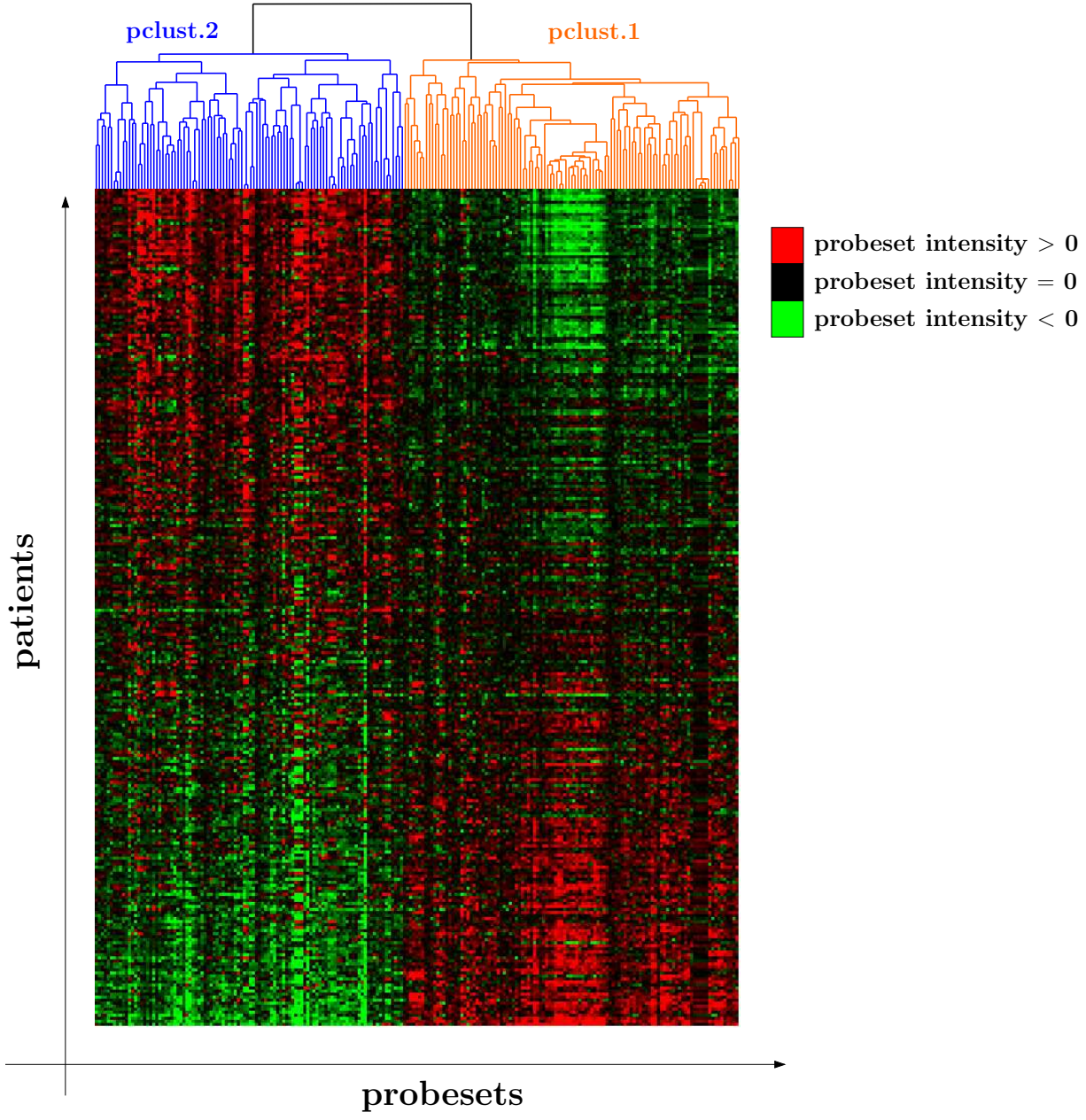
- Maximization w.r.t. $\boldsymbol{\beta}$ using Newton-Raphson algorithm [Collett, 2003].

Performance Estimate

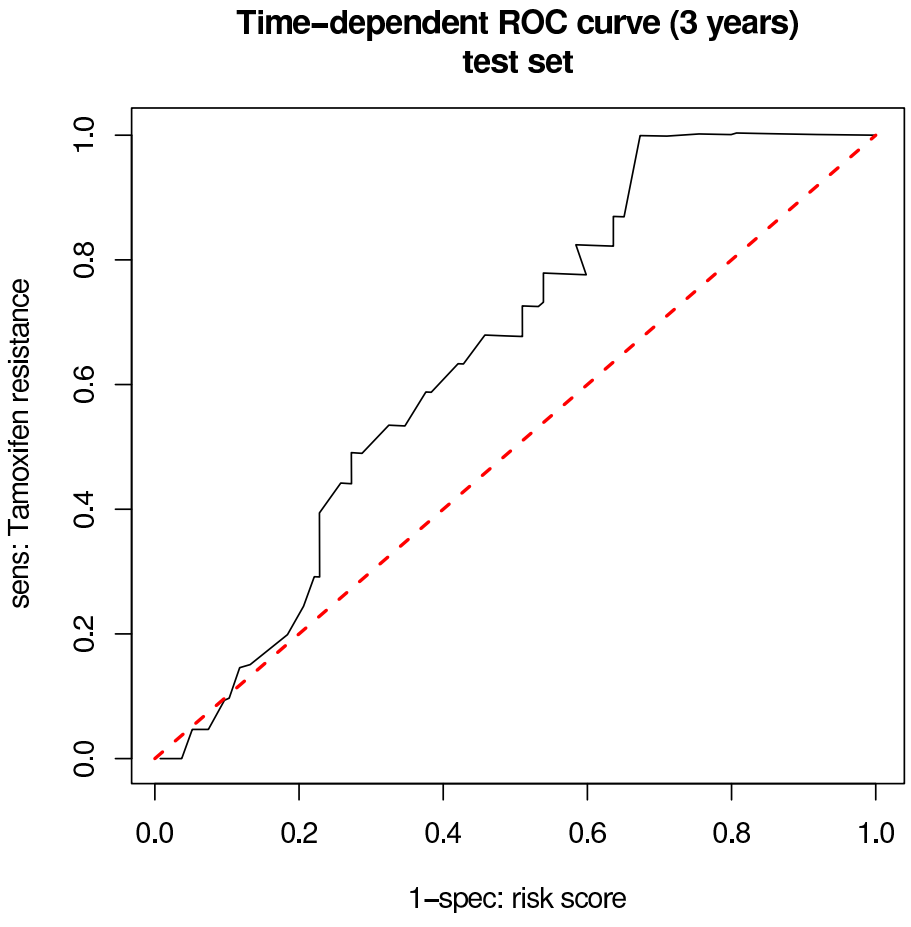


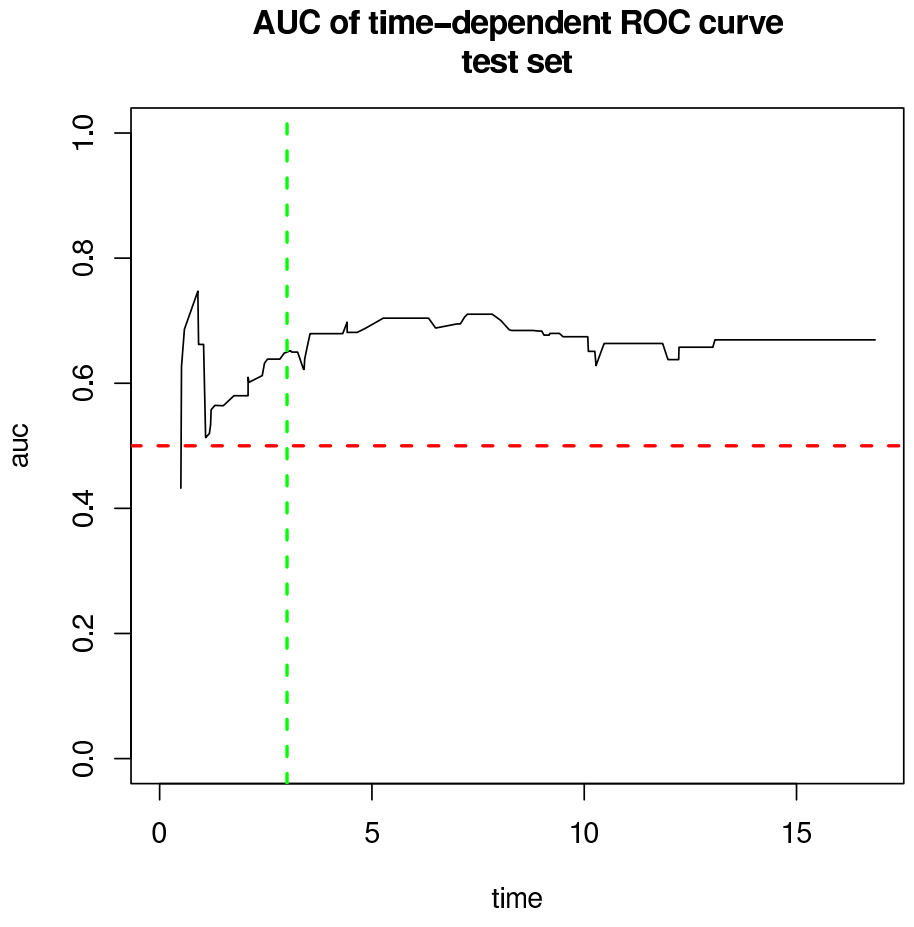


Gene Expression Visualization



TD ROC Curve on Test Set





Bioinformatics Software

- **R** is a widely used open source language and environment for statistical computing and graphics
  - Software and documentation are available from `http://www.r-project.org`
- **Bioconductor** is an open source and open development software project for the analysis and comprehension of genomic data
  - Software and documentation are available from `http://www.bioconductor.org`